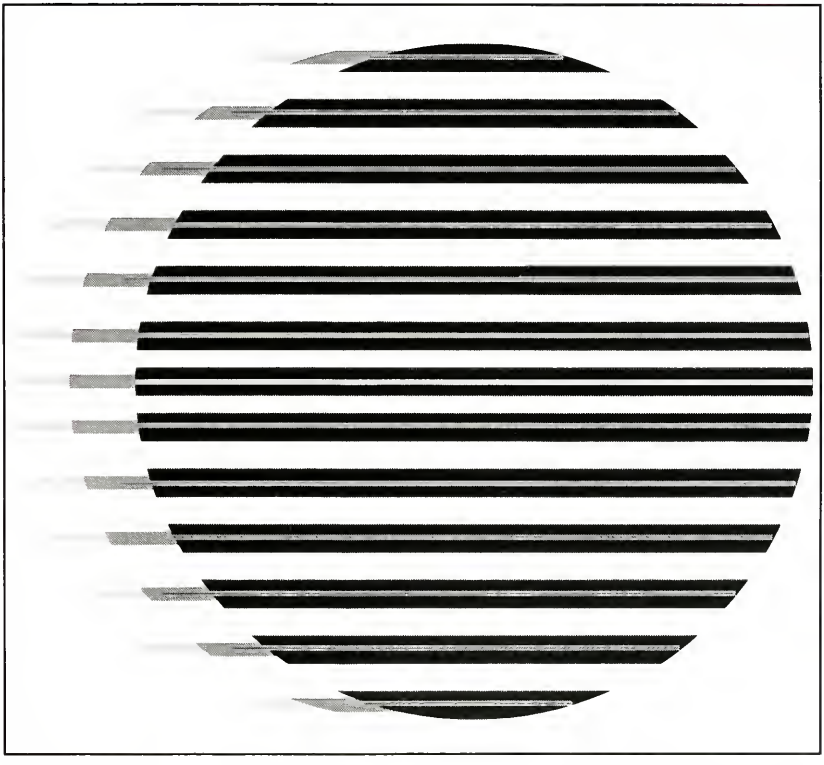
IASSIST

Q U A R T E R L Y

VOLUME 24 Fall 2000 NUMBER 3



Printed in the USA

IASSIST QUARTERLY

The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors:

The **QUARTERLY** is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to jtstratford@ucdavis.edu. Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979*. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Title: Newsletter - International Association for Social Science Information Service and Technology

ISSN - United States: 0739-1137

CONTENTS

Volume 24 **Number 3** **Fall 2000**

FEATURES

University Information System RUSSIA: Scientific and Social Challenge

*by Tatyana Yudina**

An appropriate information base is the main challenge for research and education in social and human sciences in Russia, especially in distant areas. Due to an information shortage, the general level of teaching and applied investigations is decreasing - university professors are unable to recommend as obligatory for study recently published books and periodicals: funding for book purchases is poor. As the Ministry of Science of RF (the Russian Federation) recently reported, only 67 scientific journals are available for 10,000 investigators in Russia (408 - in Great Britain, 186 - in the USA). Official government documents and reports, state statistics are also not available for educators and investigators. Due to the lack of public domain state statistics, the new research methods based on processing of large sets of numerical data are not developed in Russia. In the current situation, Internet-based collective resource is not only the most rational but the only possible approach to arrange the information supply and build the research base for investigations and advanced education in Russia.

The Moscow State University (MSU) Research Computing Center and non-commercial organization Center for Information Research since 1994 have been working to meet the challenge and develop the University Information System RUSSIA (UIS RUSSIA). In January 2000 the UIS RUSSIA (www.cir.ru) started operating on regular basis as a collective information base providing free access to all Russian universities. During 2000-2002 the UIS RUSSIA will compose an appropriate resource base for full-scale investigations in main human and social sciences. The Internet access ensures equal opportunities to researchers and educators in all regions of Russia. The universities of the former Soviet Union (FSU) countries are also granted free access.

In 2001 the UIS RUSSIA is planned to be available to local public libraries in Russia and FSU countries.

The current version of the UIS RUSSIA includes the data and documents' sources recommended as the first priority collections by the Center for Sociological Research and Economic Faculty of Moscow State University:

- official data and documents (laws, presidential decrees and directives, governmental enactments, acts and regulations) since 1991;
- stenogramms (daily records) of State Duma of Federal Assembly of RF from 1996;
- Goscomstat of RF data (all available in electronic format);
- election statistics of both federal and local levels since 1993, provided by Central Election Commission of RF;
- mass media sources (8 newspapers and 2 information agencies);
- databases, publications and reports of leading analytical centers;
- reference data on the Russian political system (brief history, prerogatives, structure and personnel of federal institutions, political parties, churches, etc.);
- extended reference information on the components of the Russian Federation.

All data collections are obtained for free from official holders/producers under legal agreements with Research Computing Center of MSU. The provision to process the information, integrate the results into the UIS RUSSIA and provide access to all universities of RF makes the UIS RUSSIA a valuable resource for full-scale socially relevant investigations.

Information update

Full text documents - official data and documents, stenogramms, mass media electronic versions are received electronically on daily basis, bulletins and analytical reports - on weekly or monthly basis (upon publication).

New full text collections will be added in 2000 :

- international agreements, signed by RF since 1991, international agreements signed by USSR,
- Constitutional Court of RF, Supreme Court of RF, Arbitrary Court of FR, decisions,
- Commonwealth of Independent States countries multilateral and bilateral agreements.
- local mass media sources.

All the documents are automatically processed – metadata created, classified, indexed, annotated and integrated into the UIS RUSSIA. The NLP technology provides for 20 Mb (equivalent of up to 10,000 pages) processed daily.

Appropriate retrospective coverage of each source will be realized during 2000.

Statistical data - Numeric data are the mostly used resource for social research. State statistics are the basis for socially-relevant investigations and sound recommendations for decision-makers. The UIS RUSSIA legally obtains, stores and updates collections from State Statistical Committee of RF. Under discussion are agreements with other main statistics-producing government institutions.

The Goscomstat of RF collections are received upon publication - on monthly, quarterly or annually basis. Statistical data are received in .doc format (digital versions of publication). To make the data available for Internet search, the UIS RUSSIA specialists convert the data into HTML format; and as a next step - into Excel spreadsheet format to make the data usable for secondary analysis. Currently available are the following data collections :

- Russian Annual Statistical Report, 1999,
- Industry of Russia in 1999,
- Regions of Russia in 1999,
- National Accounts of RF, 1991-1999,
- Environment Protection in Russia in 1999,
- Finance in Russia in 1999,
- Prices in Russia in 1999,
- Russia and Commonwealth of Independent States Countries in 1999,
- Russian Annual Demographic Report, 1999.

The data collections are also indexed to make them searchable using the UIS RUSSIA Thesaurus.

In 2000, all other Goskomstat of RF collections will be added. For 2000-2001, there are also plans to obtain and integrate the data maintained by the Centrobank of RF, the Ministry of Finance of RF, Goscominyushestvo of RF, the Ministry of Labor of RF, other ministries, committees and agencies of RF, regional statistics of components of RF, international organizations measurements, and the databases created under the foreign grants. The numeric data collections are complemented by methodological notes.

Part of the Statistics of RF bloc are analytical reports prepared by leading "think tanks" in Russia - Russian-European Center for Economic Policy, Bureau of Economic Analysis, Fond for Population Sentiments Index research, etc. Reports of main Russian and foreign foundations' grantees are included.

The documents are also indexed to make the analytical materials retrievable by Thesaurus-based cross-search.

Electoral statistics are received shortly after elections. The current version stores all general elections results since 1993, and local election results. The data are converted into Excel spreadsheet format and may be analyzed using standard software packages like Statistika, SPSS, SAS, etc. Electoral statistics is region-tailored and displayed in a map format.

NLP technology

To process and integrate large scope of electronic documents the technology of Automatic Linguistic Text Processing (ALTP) is realized under the project.

The ALTP performs:

- processing of electronic text corpora in main formats (ASCII, HTML, MS Word) in Windows and operating as DLL;
- morphological analysis of Russian texts;
- terms' recognition/disambiguation;
- thematic analysis - event categorization, indexing, annotation/summarization;
- download of results on Oracle database server.

The main instrument of the technology is the *Thesaurus on Contemporary Russia (Thesaurus)*, created under the UIS RUSSIA project. In its current version it incorporates 18,500 concepts/descriptors, includes 6,500 geographic names, 39,000 synonyms, 70,000 relations between concepts, 200,000 inherited relations. The tool assists in detecting of main and subordinate topics in a document as a result of analysis of macroconcepts and relations between them. Macroconcepts are modeled by groups of concepts semantically related in Thesaurus. Thematic representation provides for evaluation of weigh of each term in a text and performs event categorization and annotating/ summarization of a document. The Thesaurus enables to determine up to 90 - 95 % of terms.

The technology provides for up to 20 Mb of electronic texts to be processed on each Pentium200 PC and integrated into the University Information System RUSSIA daily.

The technology was evaluated by experts from NIST and DARPA in 1996 under the TextRetrievalConference-6 program and SummarizationConference in 1997. The results are among the best in a group of 14 participants.

The ALPT ensures advanced search instruments.

Search engine

Being initially designed to serve scientific needs the UIS RUSSIA provides for advanced search instruments: in

addition to traditional tools it includes value-added elements - the System of Subject Headings and Thesaurus on Contemporary Life in Russia. The System of Subject Headings consists of 200 topics (rubriks). All full text sources are filtered and event categorized according to System of Subject Headings. The Congressional Research Service, LC Legislative Indexing Vocabulary-based search is also available.

The UIS RUSSIA provides for research assistance, user services, metadata and annotation browsing, and thesaurus based query refinement. User-tailored automatic information update is realized.

Technical base: Architecture
The UIS RUSSIA operates at the Research Computing Center of MSU server. Mirror sites will be maintained in Novosibirsk and St. Petersburg to ensure more reliable access for universities in the Northern part of RF, Siberia and Far East.

Analytical bloc
Educational activity in advanced research methods has been started. Main research institutions were contacted and several software programs, Workbench of Sociologist, Workbench of Economist, etc., are presented by the authors. Special training class is arranged by the Research Computing Center of MSU, where the analytical software is downloaded and made available for university faculty. Special course is scheduled, it includes lectures analyzing main approaches to computer-based investigations and demonstration of working models, teaching and training in basic and advanced technique of social quantitative analysis. The main idea is to make available for investigators and educators the sound projects, to store the research results costly in both financial and human terms and to preserve them for future use. Consultations of authors of the software will be available for the faculty ready to use the programs in educational courses and investigations. Socially-relevant projects using the UIS RUSSIA stuff will be initiated. Not only MSU faculty is informed and invited but other universities of Moscow and regional universities.

Bilingual search instruments
The UIS RUSSIA has been initially designed as part of the international information structure to serve not only Russian researchers but also foreign specialists on Russia and general public. To meet the challenge, a special complex of the bilingual searching tools is being developed. The prototype of the bilingual complex is ready for testing and evaluation by a team of Russian American specialists. Funding to evaluate the bilingual search tools is currently being sought.

The NLP technology and developed bilingual complex will produce an annotation in English on each Russian document. This accomplishment widens the UIS RUSSIA audience, helping foreign public to open Russia and foreign specialists to investigate Russia.

Work is underway on the Global Information Service (GILS)-profile to provide for the UIS RUSSIA integration into the world information space assisting Russian specialists in international cooperation in economic, social, political, human research.

Russian universities network
The UIS RUSSIA has been designed as a base for inter-university cooperation in consorted and rational efforts to build a networked collective information infrastructure. The regional universities may actively participate. Currently up to 50 local universities are technically and technologically equipped to take part in the cooperation. The Open Society Institute (Soros Fund)-funded "Russian Universities' Internet Centers" program provided the regional universities with the hardware-software platform compatible with that of the UIS RUSSIA. The NLP technology developed under the UIS RUSSIA project may be passed on for free to the regional universities to enable them to develop information systems of their own on local resources, integrated into the UIS RUSSIA. The local stuff is important to make the social analysis relevant.

In this respect the UIS RUSSIA is close to the American universities' Internet2 initiative directed to rationally build Internet-based far-reaching educational and research network.

From the very beginning the UIS RUSSIA project was developed in cooperation with the Michigan Inter-university Consortium for Political and Social Research and European Consortium for Social Research. Specialists of both structures visited Russia and have been of help to the Russian researchers.

Program to become self-supportive
Maintenance of the system on self-supporting basis is a challenge of the project. The experiences of information structures in the USA, Europe and other countries' have been analyzed, and main elements of financial activity of those organizations will be realized - institutional membership with annual dues for foreign universities and other organizations. Preliminary discussions with American and European colleagues - university professors, think tank specialists, government analysts, journalists prove that this way is the most acceptable for them as well. The dues will create a relatively stable and predictable financial base and allow the program to engage in long-range policy to develop the information resource and provide access for free to the high education institutions in Russia.

The team
The UIS RUSSIA project began 1994. The key specialists have worked together since that time. The team includes 20 specialists from the Research Computing Center, other faculties of Moscow State University, academic institutions and other universities of Moscow. Several specialists are invited for half-time job and consultations. A group of American consultants provide their expertise of the project on the permanent basis.

Since 1994 the project has been supported by grants from Russian Fund for Basic Research, Russian Humanitarian Scientific Fund, Ministry of Science and Technologies of RF "Informatization of Russia" program, MacArthur Foundation, USA, Ford Foundation, USA.

* Tatyana Yudina, Ph.D., Leading researcher of Moscow State University Research Computing Center, Director of University Information System RUSSIA project.
yudina@mail.cir.ru

The Social Science Electronic Data Library: Serving the Needs of Data Librarians and Users

by Michael Carley &. Josefina J. Card *

The last decade has witnessed enormous strides in two areas: first, the development of numerous social science data sets of high quality; and, second, the development of the computing hardware and software capability and infrastructure needed to locate and analyze these data sets for minimal cost and to communicate data findings in interesting and easy-to-understand fashion. Hand in hand with these advances in data development have come technological advances which allow social science research and teaching laboratories, with the hardware and software needed to analyze the best data in a given field, to be set up with ease by an academic department or even by an individual professor. Additionally, sophisticated data analysis software packages formerly available only for mainframe computers have become available for microcomputers at a much reduced cost. Taken together, these developments make it possible for academic departments, research institutes, and government offices of all sizes and levels of financial resources to access and analyze exemplary data sets for research, teaching, and program- and policy-development purposes.

Data archives, in both the private and public sectors, allow easy and open access to many hundreds of the best health and social science data sets covering a broad range of topics, study populations, and making use of a variety of research designs. The data available from many of these archives are clean and the documentation user-friendly. For these and other reasons, researchers and instructors who are considering the use of secondary data welcome the functions served by a well designed data archive. This data is used to:

- conduct secondary analyses of outstanding data sets to serve the needs of policy, practice, or basic research;
- perform meta analyses based on access to multiple original raw data sets;
- prepare research proposals on various issues;
- write publications comparing and contrasting results from related data sets;
- prepare masters theses and doctoral dissertations;
- produce classroom materials for teaching substantive, methodological, and statistical concepts from real-world data.

In this paper, we will discuss three areas of major concern for data librarians and data users wishing to obtain and use data for secondary analysis: data quality, format, and dissemination. We review and contrast the issues and concerns of data librarians and data users, outlining areas of similarity and difference. We will explore how one large data collection, the Social Science Electronic Data Library (SSEDL), compiled over the last 17 years by Sociometrics Corporation, has addressed each of these issues and the conflicts and problems that arose during that process. Finally, we peer into the future, assessing how data providers can bridge knowledge gaps via recent technological advances.

The Social Science Electronic Data Library (SSEDL)

The Sociometrics Social Science Electronic Data Library is a premium health and social science resource that consists of seven topically focused data archives. With over 300 data sets from 200 different studies comprising seven topically-focused collections, it is a unique source of high quality health and social science data and documentation for researchers, educators, students, and policy analysts. The Electronic Data Library was made available in 1999 on a set of CD-ROMs and includes an online membership with free access to datasets for downloading by members.

The Collections:

The **Data Archive on Adolescent Pregnancy and Pregnancy Prevention** (DAAPPP) was established by the US Office of Population Affairs (OPA) in 1982 as the repository for the best social science data on the incidence, prevalence, antecedents and consequences of teenage pregnancy and family planning. In 1994, the scope of DAAPPP was expanded to include studies that focus more broadly on adolescent sexual health issues, thereby including studies examining behavioral factors related to sexually transmitted diseases (STDs) in addition to pregnancy. DAAPPP currently holds data from over 150 premiere studies (many of them longitudinal) on sexuality, health, and adolescence.

State-of-the-art research data on the American family are

available through the **American Family Data Archive** (AFDA). AFDA, funded by the National Institute for Child Health and Human Development, contains data and documentation from 20 nationally recognized studies on important issues relating to American family life, demographics, and family patterns. Among the topics covered are educational, economic, health, social, and psychological indicators, child welfare, family violence, marriage, divorce, child care and child custody.

The **AIDS/STD Data Archive** (AIDS) consists of original research data and instruments from 11 premier studies on AIDS/HIV and other sexually transmitted diseases (STDs). The collection was established with funding from the National Institute of Child Health and Human Development (NICHD. Included data sets address the following topics: the incidence and prevalence of specific sexual behaviors (including abstinence, vaginal and anal intercourse, oral-genital sexual activity, masturbation); contraceptive and STD-preventive behavior; attitudes and beliefs regarding sexual behavior and methods of contraception and STD prophylaxis; AIDS/HIV knowledge, attitudes, behavior, and serostatus; current and past episodes of gonorrhea, syphilis, chlamydia, and other STDs; and high-risk behavior, including alcohol/drug use and prostitution.

The **Maternal Drug Abuse Archive** (MDA) brings together seven state-of-the-art research databases on maternal alcohol and drug abuse. Funded by the National Institute on Drug Abuse, the collection includes data on the following topics: the prevalence of drug use among pregnant women and women of childbearing age; demographic characteristics of pregnant drug users; types and patterns of illicit drug use; social, psychological and economic antecedents of pre- and perinatal drug abuse; the effects of pre- and perinatal substance use on pregnancy complications and neonatal status; and the effects of fetal alcohol and drug exposure on children's physical, neurobehavioral, psychological and social development.

The **Data Archive of Social Research on Aging** (DASRA) was assembled with the support of a grant from the National Institute on Aging. DASRA contains data and documentation from three very large nationally recognized studies. These three studies covered a variety of topics including functional status and impairment, living arrangements, caregiving and social support, health attitudes, retirement income and plans, mortality, health, financial resources and assets, expenditures, cognitive ability, medical conditions, housing, health insurance, and personal characteristics

The **Research Archive on Disability in the United States** (RADIUS) was funded by the National Center for Medical Rehabilitation Research (NCMRR) within the National Institute for Child Health and Human Development (NICHD). The purpose of the project is to facilitate access to the best data sets on the prevalence, incidence, correlates, and consequences of disability in the U.S. The heart of the archive is a collection of 19 studies that address the topic of disability. These data sets permit analyses on topics such as: the incidence and prevalence of specific diseases, disorders, and impairments, including deficits of cognition, emotion, physiology, and anatomical structure; functional limitations across a variety of specific organ systems; disabilities in relation to major life roles and activities, such as work, parenting, education, and recreation; societal limitations including physical, attitudinal, and economical barriers that restrict full participation in society; psychosocial and interpersonal factors such as coping with stress, sexuality, feelings of control and productivity, quality of life, and family relations and support; health care and rehabilitation issues such as medical costs, coverage, service utilization, use of orthotic, prosthetic, assistive devices, effectiveness of rehabilitation; as well as a variety of basic demographic factors on respondents such as age, race, sex, income, occupation, marital status, family size, and living arrangements.

To facilitate access to the best contextual data, Sociometrics has developed a **Contextual Data Archive**. By contextual data we mean data that describe the population, social, and economic characteristics of geographic areas, from census tracts to states, in which people reside or work. The contextual data archive consists of a series of files, each organized around a different geographic unit of analysis (such as census tracts, school districts, counties, states, etc). Each file contains variables drawn from various sources, but having one common geographic unit of analysis. Support for this project was provided by the National Institute of Child Health and Human Development.

Data Quality

Both data librarians and data users have an abiding interest in the availability of high quality digital data. The librarian must put her/his limited resources to the most efficient and effective use possible. The resources we mean here are not only financial assets such as approved budgets, but also material and human capital such as shelf space and the person-hours of those who must purchase, assemble, and maintain various data collections. Because all of these resources are limited and precious, data librarians must make wise choices as to how to prioritize their use in order to achieve the highest quality collection of data for the users at their institutions.

Data users are also concerned about having data of the highest possible quality. The users of digital data wish to make the best possible contribution to the body of knowledge in their field. That contribution is placed in jeopardy if the data used is of questionable quality. Data gathered via poor research design, or via a good design

poorly executed are of little use in advancing knowledge. Researchers using such data risk not only making a tainted contribution, but also of generating criticism from colleagues and associates who recognize the problems or limitations of the data being used.

The research staff at Sociometrics recognized these needs when compiling the seven data archives in the Social Science Electronic Data Library. It was determined that each archive would be a 'best of the lot' collection, accepting only the best data available in each of the seven topic areas. To accomplish this, we formed National Advisory Panels of research scientists who were experts in both the substantive content of the particular archive and the research methods commonly applied in that field. The panel, usually consisting of six members, was asked to evaluate candidate data sets on the following five criteria:

- **Technical quality**: among the factors to be considered are high response rates, low attrition rates, use of reliable and valid measures, and sound sampling and design elements.
- **Substantive importance to the field**: factors include the potential to address contemporary issues, to break new ground, and to replicate or confirm important findings.
- **Program or policy relevance**: the ability of the data set to answer applied questions on how to improve public policy or shape intervention programs;
- **Potential for secondary analysis**, including: *Scope of sample* - The broader or more diverse the scope of the sample, the greater the potential of the data for generalization. *Size of sample* - Sample size is always an important consideration. This is even more true for data intended for secondary analysis: sample sizes adequate to support the originally intended analysis may be too small to support other analyses, especially if the new analyses focus on data cells that have a very low proportion of cases. *Breadth of variables and constructs covered* - The potential for secondary analysis is directly related to the breadth of variables measured in the data set. The more numerous and diverse the set of variables, the more possibilities there are for new or expanded analyses.
- **Disciplinary balance**: An archive should attempt to be representative of the entire field of research. Variations in state of the art exist between different sub-areas within any discipline. Thus a somewhat flexible standard (as measured by the other criteria above) should be used to ensure that all major areas of the discipline are represented in the archive as a whole.

In order to perform these evaluations, we provided each panel member with briefing materials consisting of a 2-4 page description of the data source, which covered: the purpose of the study; methods (including sampling design, periodicity, unit of analysis, response rates, and attrition); content (description of variables covered, number of variables, and topics covered); limitations; sponsorship; and a bibliography. In addition, we provided copies of original peer-reviewed publications for each data set, which allowed the panel members to review issues we may have not addressed in our briefing documents. Panel members were encouraged to suggest additional data sets for consideration, and have often done so. Panel members did not vote yes or no on each data set, but rather rated each with a 'priority score' from 1-10. Only those data sets receiving an average score of 7 or above were accepted, and higher priority for archiving was given to those receiving higher scores.

In addition to pre-screening the data sets, archivists for the SSEDL data sets perform several other tasks designed to ensure data quality. We review the data thoroughly, checking to make sure that all variable and value labels are included and are sufficiently descriptive. We check the data for internal consistency and completeness, scanning in particular for variables with an excessive number of missing or out-of-range values. We also perform random checks verifying that the skip logic in the original instrument was followed and that the variables are consistent in relation to one another (no variables describing a female as a father or brother or a male as a mother or sister, etc). Finally, we produce a user's guide to the machine-readable files and documentation which notes any remaining limitations or inconsistencies.

Those archiving digital data face many challenges, not the least of which are the limitations on their own, as well as the users' time and resources. Consequently, different data archivists take a variety of approaches to address these challenges. Our 'best of the lot' approach emphasizes quality over quantity. This means that our data collections, while not as large as some of those from other sources, are of higher overall quality and are better documented than is the industry average. This approach limits the size of our collections, but contributes to their popularity among researchers for their high quality and ease of use.

Formats

The needs of data users and librarians diverge somewhat when it comes to format preferences for digital data collections. Users look for data in the most easily accessible form, while librarians must be concerned with the big picture, and look for collections that serve the needs of as many users as possible, both in the present and the future. The format(s) in which data are provided also have an impact on the role the librarian will take in the data distribution process, which could vary from that of a facilitator to an active gatekeeper. The challenge for data

providers is to address both the *preservation* needs of the librarian and the *ease of use* needs of data users.

Distribution Media. Librarians must conserve their many precious resources, including both financial resources and shelf space. However, they also desire that data collections be in a form that is not easily lost, damaged, or misused by careless users. Therefore, a data collection should be durable, and in a format that is not likely to change rapidly with evolving technology. CD-ROM technology meets these requirements. CDs are more durable than diskettes and do not have the associated (at least perceived) transient qualities of internet sites. Data made available on a CD-ROM are not likely to be easily lost as library staff can, should they choose to, maintain tight control over them or copy their contents to a central repository for safekeeping. Diskettes are more likely to become corrupted and internet sites often are revised and require more constant updating on the part of both the data provider and the librarian to keep all links accurate and up to date.

Data users, on the other hand, prefer that the data are made available in the simplest, easiest to access format possible. However data are made available, it must be transferred to the computer where the user will actually be working. With desktop computer speed and hard drive space increasing exponentially, users often prefer that data be available for copying to their own system, rather than residing at a central repository. Internet downloads may be preferred to CD-ROMs as the data can be copied to the user's own computer, then manipulated and transformed as necessary.

To best accommodate these divergent needs, we found it necessary to make our data sets available in both CD-ROM format and via our internet web site. SSEDL Volume 1 is distributed via 17 CD-ROMs, along with accompanying support material. Additionally, each purchasing institution is given free web access to all of the data sets, as well as to new data that have not yet been added to the CD collection. By allowing the users to download the data sets or use the CD-ROMs, we were able to provide both the user and the data librarian with flexibility in both data format and data access.

Analytic Software. A user who wishes to use data on a particular topic would be best served by data that can be retrieved quickly and effortlessly with a variety of software. Given the wide variety of statistical software available to users in different fields, this can be a challenge. Users should have the capacity to use the software of their choice, and the ability to access the data quickly with that software. At the same time, data providers must understand that software currently popular may change or become outdated, making files created from these programs unusable or at the least cumbersome and inefficient to use.

Most data collections have taken one of two approaches to this problem. First, the data provider may distribute *raw data* with a *codebook*. The raw data is typically stored in an ascii file which is simply useless text (numbers) without the codebook. The codebook provides the user with the location of specific variables and cases within the raw data file. The advantage to this approach is that it addresses the issue of durability well. Users can access the data by writing a program using the statistical software package of their choice, inserting the variable and case locations given in the codebook to access the data. Changes in software applications do not affect data distributed in this method, as users write their own program with the language in which they have expertise. However, writing the programs to read the raw data can be a time consuming process, causing users to waste much of their resources on mundane tasks. In addition, such writing is prone to error; one misplaced character can cause much of the data to be written incorrectly.

Other data providers address these issues by distributing the data in a pre-packaged format using one of the most popular statistical software packages (typically SPSS or SAS). By distributing these formatted files (usually either complete system files or portable files), the user can access the data directly simply by opening the files in the appropriate software. When portable files are used, the data can be used with different versions of the same software package (either earlier versus later versions or versions for different operating systems) or in some limited cases, in other popular statistical packages. The advantage to this method is clear: quick, easy access to data. The disadvantage is that this approach cannot possibly be flexible enough to address all user needs. Some users wish to access the data with a software package that is not among the most popular. Also, data distributed in this method can become unusable when software packages radically change their formats. Data made available in the most recent format could be inaccessible within just a few years.

To address the limitations in each of the above approaches, data sets in SSEDL are distributed with raw data files and machine-readable set-up statements for use with both SPSS and SAS statistical software. These set-up statements provide for the best of both worlds: ease of use, combined with flexibility. Users use these syntax files to create the system or portable files in whichever software they are using. For those users who are use software other than SPSS or SAS, the set up statements serve essentially the same purpose as the codebook described above. Because the set-up files are machine-readable, they can often be converted for use with other software with a minimal investment of time. Should the syntax requirements of SPSS or SAS change radically, these files would serve also accomplish this goal.

In addition to the above files, SSEDL data sets also are distributed with a machine-readable SPSS data dictionary file and an SPSS frequency and statistics file. These aid the user in making sure that they have created their system or portable files correctly. Users can compare the statistics in the frequency file to their own, thereby preventing mistakes in data analysis. Both the frequency and dictionary files are also useful in reviewing the contents of the data set. Each data set is also accompanied by a printed User's Guide (provided in machine-readable form, in addition to printed form, for the more recent archives) comprised of a standard set of sections and subsections. The provision of standard machine-readable and printed documentation assists users in familiarizing themselves with the Sociometrics data sets. Once a user has worked with one Sociometrics-packaged data set, it is easy for him or her to work with any of the others. The original instrument and codebook are offered as optional, supplementary documentation for each data set, when available. For the more recent archives, the original instrument is distributed in machine-readable form along with the data, as a set of graphics files (page images).

Search and Retrieval Software. As data sets get larger, both in the number and scope of variables covered and in the number of cases, users are faced with an increasingly overwhelming task of reviewing which parts of a study are necessary and appropriate to their needs. Often, users will begin work with a data set containing over 5,000 variables (and often several thousand cases) only to find that their interests only require 30-40 of those variables. It is important for users to be able to quickly sort though the variable list and find those of interest. Given current (though perhaps temporary) limitations in speed and disk space, users also need to be able to reduce the large data set into one with only those variables needed for analysis.

To address this need, Sociometrics staff developed powerful search & retrieval software which now accompanies each data archive. This software allows a user to search an entire topically-focused collection, a customized group of data sets created explicitly for a given user, or a single data set; to identify variables of interest across this designated search space and to save located variables as a search set. Users can conduct: (1) full-text keyword searches, including variable names, words in variable labels (question descriptors), and words in value labels (response descriptors); (2) searches by assigned topic and type codes; and (3) searches by study name or assigned data set number. Standard Boolean operators (i.e., "and," "or," "not") can be used to combine search sets. Alongside this software, we provide data extract software which allows users of CD-ROM versions of archived data sets to create customized SPSS or SAS program files containing only those variables of interest to them. This capability permits analyses of subsets of large data sets to be conducted quickly (with rapid turn-around) on most microcomputers. It also saves users significant program development time writing and re-writing SPSS and SAS program statements to define variables used in a given analysis.

Technical Support

The Role of the Data Librarian. The role of the data librarian in this process varies a great deal among institutions. In some cases, the librarian serves as an expert gatekeeper to the data, allowing access to users as s/he deems appropriate and answering a wide range of questions users may have. Others may serve a minimal role, simply providing access to the data and support materials and little else. Librarians also vary in their level of statistical knowledge, as well as their expertise in the various topics that may be covered by the data in their collections.

Our approach to this issue was again to provide the greatest amount of flexibility in the collection as possible. While some data librarians do take on something close to a gatekeeper role, we chose to allow for those who had neither the time nor the expertise to do so. Librarians need to be fully informed, not on all of the topics included in their data collections, but on the process by which they can aid users in finding data of interest to them. Given the wide variety of topics covered in SSEDL, one could never expect librarians to provide users with all of the help they may require. To this end, the Social Science Electronic Data Library includes a variety of tools to facilitate this process. A user's manual and contents manual detail the data sets included in the collection, and a quick start guide offers advice on how to implement the software and the knowledge needed to use the data sets. Most importantly, a 'Guide to the Social Science Electronic Data Library' CD is provided with each collection. This CD takes the librarian through the process of using the data library using a brief step by step tutorial.

SSEDL's Research Support Group. In addition to the support given to the data librarian, users have direct access to help from Sociometrics' archiving and scientific staff through our Research Support Group (RSG). The Research Support Group consists of Ph.D. and Masters level social scientists who provide free technical assistance for users who have questions about accessing or using our data sets. In addition, the RSG occasionally performs consultant work such as the creation of customized data set extracts; user-defined statistical tables and analyses; data archiving, management and analysis services; customized CD-ROMs; and training workshops. These services greatly aid users with limited expertise or resources with which to conduct their own analyses.

Dissemination

The manner and methods by which digital data are disseminated by data providers and eventually by data librarians are crucial to the usability of such data. Users

must be made aware of the availability of data that meets their needs. However, with today's rapidly expanding technologies, the problem of 'information overload' is a crucial one. Even within our own collections, users can become overwhelmed with the sheer amount of information available to them. If these issues are not handled properly, they can inhibit the users' ability to locate and make use of the most appropriate data. Data providers must do what they can to ensure that users have the capability to quickly and easily locate the data sets and even the particular variables their topic of interest requires.

While the search and retrieval software made available for each individual data archive helps users find variables within a study they have already chosen, it does not help when users have not yet selected the study that meets their needs. To address this issue, we created a search mechanism for use on our internet site which is cross archive. This allows for users to search by keyword(s), or designated variable topic or type, for variables of interest in all of the SSEDL data sets. Users can search for words in variable labels (question descriptors) or in value labels (response descriptors). In addition, we include a brief abstract of each study, and users can search for keywords within those abstracts. We chose to make this software available to the general public as well as users on our internet site, in order to allow researchers at non-purchasing institutions the opportunity to find data sets of interest and order them individually.

Data librarians must also make an effort to make help users become aware of available data collections. In order to facilitate this process, we provided not only the above mentioned guide to SSEDL on CD-ROM, we provided informational flyers and brochures to help the librarian make potential users aware of the availability of our data collections. In addition, the Research Support Group provides both librarians and users ongoing advice as to how to find data sets of interest in our collections.

Looking to the Future: New Technologies, New Audiences

The value of any data collection is in part predicated upon its ability to address issues of the day. Therefore, any collection will inherently be of greater value the newer the data are that are contained within it. The preservation of historic data is clearly important, but any collection that aspires to be useful must also be kept up to date with the addition of more recent data. We will continue enlarging the content and capabilities of our data set collections. We will be adding to our current data archives as funds permit, and expanding our efforts by adding new topic areas to the collection. We expect to begin the establishment of a data archive on child well-being shortly. A feasibility study on the formation os a complementary and alternative medicine data archive has just been successfully completed.

In putting together the SSEDL, Sociometrics' staff have learned a great deal about data collection methods and ways to improve efficiency and reduce costs to researchers. Currently, we are developing a software product that will aid researchers on this aspect of the process. Sociometrics' Automated Dataset Development Software (ADDS) is an integrated software program that, when completed, will develop and document social science research studies. The program will perform the following functions: 1) Instrument generation—generate a fully formatted research instrument in print, ASCII, and other machine-readable formats. 2) Codebook generation—generate the data set documentation in a printed codebook (also in ASCII and other formats), flow chart (skip map), and data file map. 3) Data entry—provide for data entry from completed questionnaires, with simultaneous error checking. 4) Program file generation—produce a raw data file in ASCII format, and build the program statement files needed to transform the raw data file into SPSS and/or SAS system files. The software will automate tasks best done by computer, improve instrumentation and documentation by providing a complete, high-quality structure and format, and reduce the post data-collection effort of documenting a public-use data set.

Additionally, we are also building an item bank of high quality, commonly used questions, scales, and interviewing tools from the SSEDL collection. This bank will be accessible within the ADDS program to permit users to select questions to develop their own research instruments. The item bank will be filled with several thousand questionnaire items drawn from some of the leading studies in research on the American family. Using questions or scales that have been previously tested will not only improve the choice of questions, but will also lead to greater comparability between studies and over time.

In addition to keeping the Social Science Electronic Data Library current and helping researchers imrpove their methods, we hope to reach new audiences through new technologically innovative products. Bridging the 'knowledge gap' is of prime importance to those who wish to make practical contributions through social science research. Rather than limiting our efforts to trained researchers, we must reach out to other professionals, and, when possible, the lay public as well. We are beginning our efforts to reach the 'paraprofessional' audience with two new products related to the SSEDL. The U.S. Social Surveys: A Sampler of Questions and Responses, will contain searchable, edit-ready, and print-ready machine-readable versions of the demographic, behavioral, and health science instrumentsæquestionnaires, medical forms, interview protocolsæused to collect the data in SSEDL. Questionnaire items will be linked to crosstabulations with age, race/ethnicity, and gender, obtained from the linked SSEDL data archives.

Secondly, the Multivariate Interactive Data Analysis System (MIDAS), will allow online analysis of the data in SSEDL. Online data analytic procedures will include weighted and unweighted frequencies, percentiles, and measures of dispersion and central tendency, as well as two-way and *n*-way tables with measures of association, comparison of means (2-group and ANOVA) and correlations, and the calculation of complex variance estimations. Users will be able to define case subsets, recodes, or aggregations for analysis, and then produce output which can be downloaded or printed. Custom dataset downloads will also be available.

The goal of ADDS is to aid expert researchers in handling the 'front end' of the research process. Through the use of the ADDS software, researchers will be able to reduce their costs and improve the accuracy and efficiency of instrument development, data collection, input, management, and analysis. The goal of Social Surveys and MIDAS is to increase the accessibility of the data to those who are not competent in the sophisticated statistical software packages such as SPSS or SAS. These products will help the 'paraprofessional'--people with college degrees who are not necessarily trained in complex data analysis— avail themselves of exemplary social science data. They will provide a basic introduction to social science methodologies as well, and will be linked to the SSEDL data sets for those who wish to progress to the next stage of data analysis (e.g., advanced undergraduate students).

In sum, the historic progression of SSEDL has been to expand the definition and purpose of a data archive. SSEDL staff have worked to enhance archiving methods to make data easily accessible to researchers. Our advances in this field have helped to make the research process more efficient, especially for those conducting secondary analysis. ADDS will improve the process for primary research as well. Non-researchers will be introduced to data analysis through the new products, Social Surveys and MIDAS. Together, these products will allow us to extract the greatest possible value from our research dollars as data will be used in as many ways as are feasible and by a much wider audience.

* Michael Carley and Josefina J. Card, Sociometrics Corporation, Contact Name and Address: Josefina J. Card Sociometrics Corporation, 170 State St. Suite 260, Los Altos CA 94022, (650) 949-3282, ext. 211, FAX (650) 949-3299, jjcard@socio.com

Accessing Indian Numeric and Statistical Data: a critical study of the Suprastructure and Infrastructure in India

*by Dr. Jagtar Singh & H. P. S. Kalra**

Abstract

The use of numeric and statistical data for macro and micro level decision-making, development planning, and socio-economic research has always been critical for governments, international organisations and society at large. While the developments in information and communication technologies (ICTs) have paved way for timely access to validated numeric data on the one hand, these have also posed many challenges before library and information professionals to exploit the opportunities made available by the ICTs to manage and disseminate the numeric and statistical data efficiently and effectively.

In India, numeric data have been published regularly, mainly by the Central and State Governments. Numeric data published by the government ministries, departments, and other agencies is largely print-based, basically brought out in the form of reports, as well as ad-hoc and regular publications. Although the technology for digital storage and dissemination of numeric data had been available for a long time, its importance seems to be realized only recently by the Government. Although the NICNET (a national network for dissemination of the government data and information) has been operational since the 1980s, a comprehensive National Policy on Dissemination of Statistical Data (NPDSP) was announced only in May 1999 by the Government of India (GoI). This paper critically evaluates the provisions of this policy and also looks at the infrastructure being made available in the form of NICNET.

Though a few efforts have been made in India (e.g. by the Reserve Bank of India, and Registrar General Office) to digitise the existing numeric and statistical data, access to the digitised data is not adequate and reliable. In fact, the conduit is available, but the content is far from satisfactory. There is a strong need for assessing user needs, enhancing their awareness, and consolidating the efforts of various ministries, departments and other source agencies to make the collection, validation, organisation and dissemination of numeric and statistical data efficient and effective. A beginning has already been made in this direction by making the Department of Statistics, Government of India responsible to serve as a nodal agency in this regard. An effort has been in this paper to raise a few issues and put forward a few suggestions to ameliorate the situation and also to enhance global community's awareness regarding the state-of-the-art in India.

Introduction

The importance of reliable and accurate numeric and statistical data has been duly recognised by scientists, social scientists, planners, and decision-makers, entrepreneurs, and governments. Now-a-days, a lot of time money and manpower is invested in collecting, analysing, and disseminating information in quantitative form. As the number and variety of data sources increase, be it an academic, industrial or government setting, the process of providing access to data gets complicated. The use of traditional methods of data collection are tedious and cumbersome. All sorts of clarifications and explanations regarding the data need to be mentioned to all data source agencies/individuals time and again. Dissemination of the data analysed (in some cases unanalysed also) to all concerned poses problems for the library and information professionals. The problems are compounded with the increasing demand of users for numeric data customised according to their needs, or in ready to use formats. Users from academic organisations, bureaucracy, government, business, industry, and non-government organisations rely heavily on numeric and statistical data for their work. Though the use of computers and storage media on the one hand has solved the problems of storing and analysing large quantities of such data, it has given rise to operational problems on the other.

With the convergence of computer and communication technologies and the emergence of networks, information handling processes have undergone a profound change. Developments in information and communication technologies (ICTs), particularly in the 1990s have changed very significantly the way we manage information, including statistical information, right from its generation to use. Now it is possible to access numeric and statistical data via the networks, particularly the Internet. In the networked environment, access to, validation, security, and updating of statistical data are some of the challenges facing the library and information professionals.

State-of-the-Art Report

In the context of developments in ICTs at the global level, the situation in India regarding the availability of and access to numeric and statistical data is not so encouraging. Numeric data are collected, analysed, validated and disseminated by ministries, departments, and agencies of the central and the state governments. These data are published as reports, ad-hoc and regular publications mainly in the printed form (e.g., Publications of Central Statistical Organisation, Human Development Reports for various years produced by the state governments of Karnataka, and Madhya Pradesh, and the Statistical Abstract of Punjab published by the Economic and Statistical Organisation, Punjab).

The Statistics Wing (SW) of the Ministry of Statistics and Programme Implementation (MoSPI), Government of India, (earlier the Department of Statistics) is the apex body for official statistical system in India. Two organisations under it, namely the National Sample Survey Organisation (NSSO) and the Central Statistical Organisation (CSO) are responsible for carrying out socio-economic surveys, field work for surveys, training, dissemination and publication, and coordination of statistical activities. Since government agencies are largely responsible for collection and dissemination of numeric data, there is generally a big time lag in the publication of numeric data. Appropriate technology was available in India for quite a long time, but has not been used optimally for quick analysis and timely dissemination of such data.

Other agencies, such as the Reserve Bank of India (RBI) and the Registrar General's Office, are also engaged in providing financial and census information respectively in statistical form. Statistical publications by various government departments and agencies are in the broad areas of national income, industry, banking and finance, census, trade, agriculture, labour, and education. In the last few years however, the Computer Centre of MoSPI has also been instrumental in making available some of the publications of NSSO in magnetic tapes, e.g., the Annual Survey of Industries 1995-96, and the Report on Energy Statistics, 1998-99. As far as the bibliographic control of publications containing statistical information is concerned, there is no single source in printed form, though efforts have been made, e.g., the Statistical System in India, 1989; the Catalogue of Government of India's Civil Publications; and the announcements in newspapers entitled 'List of new Arrivals' by the Controller of Publications, Government of India.

National Policy on Dissemination of Statistical Data

The Government of India (GoI) seems to have realized only recently the importance of disseminating the numeric information in digital form. In May 1999, The Government of India announced a comprehensive National Policy on Dissemination of Statistical Data (NPDSD) and specific guidelines for release of data. The provisions of the policy are reproduced as Annex 1 given at the end of this paper. The policy of the GoI is a welcome step in the direction of dissemination of numeric data in digital form, but certain lacunae in the policy are worth discussion. Provisions of NPDSD have been examined critically below.

Under clause (i) of the policy, it is written that the data would be available to users in the form of hard copies and magnetic media. As the policy was announced in May 1999, developments in data storage technology at that time were ahead of magnetic media. Infrastructure for data storage in optical media (CD-ROMs, DVDs) is also available with government departments and agencies. A comprehensive term to include optical media would have been better. Similarly, provisions for availability of validated data via networks, particularly the Internet, could also have been incorporated in the clause.

Clause (v) of the policy will act as a hindrance in quick and timely release of data. Under this clause, it is said the data users shall have to wait for three years after the completion of field work to get the data, in case the reports based on survey data work can not be released by the concerned government agencies earlier. Moreover, the access mechanism for data collection in such a situation has not been specified. Other clauses in the policy such as clause (vi) and clause (viii), deal with non-commercial use of data, and the Department of Statistics (DoS), GoI, acting as the nodal agency for dissemination of statistical data, respectively.

Statistics Wing (SW) in the MoSIP (earlier the DoS) has been entrusted with the responsibility of data collection from source agencies; the organisation of data and ensuring its quality; conducting studies regarding data collection and validation for each type of data source; and the dissemination of official statistics under clause (viii) of the policy, and clauses (i), (ii), (iii), and (iv) of the guidelines for release of data. Regarding secondary publications, e.g., bibliographies, indexes, and directories in electronic form, provision has been made in the guidelines for release of data (clause viii), but not much information is available on the web site of SW in MoSPI (http://www.nic.in/stat)

Though dissemination of statistical data is the focus, mechanisms have been spelt out only for release of data, and not for its dissemination. SW can act as disseminator of statistical information, to one and all only if it has a network of branch offices. It does not have any such network, and under the present provisions of the policy and guidelines, therefore, either the dissemination activity, largely print-based, will be centralised or will have to rely on some other government department/agency. Public libraries offer such a network in almost all the states and union territories in the country. Traditionally, public

libraries have been playing the role of information disseminators. Many states in India have working public library systems today. Ten out of 28 states have public library legislation. The policy could have incorporated the role of public libraries in collaboration with SW for dissemination of statistical data.

National Statistical Commission

The National Statistical Commission (NSC) was setup by the GoI in January 2000 to critically examine the shortcomings and deficiencies of the present statistical system with a view to recommending measures for a systematic revamping of the system. NSC has 12 members under the chairmanship of Dr. C. Rangarajan, the Governor, Andhra Pradesh. Terms of reference of NSC include timeliness, reliability, and adequacy of statistical system in India; dissemination of these statistics; and the coordinating mechanism using statistical information for policy making and planning. Although the commission was expected to submit its report to the Government within a period of twelve months from the date of its establishment, it has not submitted the report. (The text version of the information on the NSC downloaded from the MoSPI web site is enclosed as Annex 2.)

Infrastructure for Dissemination of Digital Data

The infrastructure for storing and disseminating the data in digital form was set up in 1977 by the Government of India under the Department of Electronics. Later, it was entrusted to the direct control of Planning Commission. Recently, recognising the growing importance of ICTs, a separate Ministry of Information Technology (MIT) was created by the GoI, with NIC and its infrastructure under the MIT. NIC has a large distributed network infrastructure, known as NICNET with its nodes in all parts of the country, including many remote areas. The conduit, i.e. NICNET, is available, but the availability of and access to the content, i.e. the statistical data, via NICNET is not easy. Effort of the NIC to provide statistical, numeric and other factual information through General Information Service Terminals of NIC (GISTNIC) is not successful as these are placed in the district commissioners' offices. Even the web site of GISTNIC does not provide users with much information (http://gist.ap.nic.in).

Although the NPDSD has been announced only in 1999, and the NSC commissioned in 2000, efforts by the concerned government agencies to provide statistical and numeric data in the digital form and via networks started earlier. Examples of these are given below. The economic and monetary data and information are available via the Reserve Bank of India (RBI) web site (http://www.rbi.org.in). RBI is the central bank of India. In addition to its main web site, it has also created special URLs for frequently accessed documents, a list of which appears as Annex 3. The documents available via the RBI web sites provide textual as well as numerical and statistical information. The Weekly Statistical Supplement, available via its web site (http://www.wss.rbi.org.in) provides economic information in numeric form under many headings (The text version of a downloaded document is enclosed as Annex 4.) Census data in the floppy discs is also available at select institutions, but its format is not user-friendly for search purposes. Brief information on census is also available from Registrar General's Office (RGO) web site (http://www.censusindia.net).

In spite of the excellent web site maintained by the RGO, not much data are available via it. Web sites of nearly all ministries and departments of GoI and their agencies exist today. A list of these sites has been compiled by Varun. A cursory look at the URLs of the web sites of various ministries, departments, and agencies reveals that NIC has created web sites for many government ministries, departments, and agencies, but adequate information is not available via the government web sites, and is not updated in some cases. In some cases, there is no email address for contacting the concerned department. Thus it becomes clear that while, with the help of the elite NIC the conduit for information dissemination has become available to data source agencies, the content (data and relevant information) available via the conduit is far from satisfactory.

Looking Towards the Future

Reforms in the telecom sector in India have been very rapid in the last few years, and with the increasing competition, prices of computers, telecommunication equipment, and services have come down heavily. More and more bodies in India are now using the computers and communication facilities and services. These are likely to increase manifold in the coming years. With such a market scenario, need for information (including numeric information) available via the computer and communication facilities, both at workplace and at home would increase considerably. Therefore, assessing user needs for statistical and numeric information and enhancing users' awareness of the existing resources and services through which they can access such information is the need of the hour.

The initiatives by the GoI, such as the announcement of NPDSD in May 1999 and setting up of NSC in January 2000 are in the right direction, but are in reverse order. The Government should have set up NSC earlier, as one of the terms of its reference is with regard to collection and dissemination of timely, reliable and adequate statistics. While the focus of NPDSD is on centralisation of statistical information system, one of the terms of reference of NSC deals with decentralisation of statistical information system. In the light of report and recommendations of NSC (whenever submitted), the

NPDSD will have to be amended or a new policy will have to be announced.

Consolidation of the efforts of various ministries, departments and other source agencies of the GoI is also needed to make the collection, validation, organisation and dissemination of numeric and statistical data efficient and effective. The SW has been entrusted with the job of coordination with other government ministries, departments, and data source agencies, and to act as a nodal agency for dissemination of statistical data (clause v of the guidelines). But nothing concrete has come out even after one and a half years of the adoption of the policy. Some guidelines and clauses of the policy act as hindrance in accessing data. In such a situation, the role of library and information professionals would be to convince the government to provide validated quality numeric information to society at large. The GoI through the SW can provide quality information by strengthening the existing infrastructure of the NIC and ensuring access to the information by making available the NICNET terminals available to people via public libraries.

Further Reading

Chidambaram, S. Siva. (1999) Access and availability of statistical information. IASLIC Bulletin, 44(3), Sep., 133-141.

Goswami, P.R. (1998) Access to socio-economic data with particular reference to India. DESIDOC Bulletin of Information Technology, 18(4), Jul, 29-38.

Goswami, P.R. (2000) Official statistical information: Indian scenario. Information Today and Tomorrow, 19(1) Jan-Mar, 11-16, 21.

India. Ministry of Statistics and Programme Implementation. (2000) Annual Report 1999-2000. [New Delhi: MoSPI]

India. National Statistical Commission. http://www.nic.in/stat (visited 10-1-2001)

Kathuria, Rajat (2000) Telecom policy reforms in India. Global Business Review, 1(2) Jul-Dec, 301-326.

National Policy on Dissemination of Statistical Data. (1999) Information Today and Tomorrow, 18(3), Jul-Sep, 16-17.

Saha, A. and Thulasi, K. (1998) Techno-commercial information on the Internet. Information Today and Tomorrow, 17(1) Jan-Mar, 6-18.

Satish Chander (1998) Access to legal information in India. DESIDOC Bulletin of Information Technology, 18(4), Jul, 21-28.

Seshagiri, N. and Reddy, C.L.M. (1997) Evolution of ethical aspects of digital information in India. International Information and Library Review, 29(2), June, 227-235.

Varun, V.K. (1998) RU on Internet? Information Today & Tomorrow, 17(4) Oct-Dec, 19-20.

Annex 1

National Policy on Dissemination of Statistical Data

i. Dissemination of official statistics in the form of reports, ad-hoc and regular publication etc. by the Central Ministries / Departments / Agencies as at present shall continue. Validated data, though published, including unit/household/establishment level data after deleting their identification particulars to maintain confidentiality should also be made available to the national and international data users in the form of hard copies and on magnetic media on payment basis;

ii. No data, which are considered by the concerned official in data source agency to be of sensitive nature and the supply of, which may be prejudicial to the interest, integrity, and security of the nation, should be supplied. The Central Government, or a state Government or the concerned Government agency, as the case may be, shall exercise its overriding prerogative to decide the degree of sensitivity of the official statistics produced by it. The data source agency will reserve the right to withhold its release altogether or to release selectively.

iii. Price of data to be supplied under (i) above should include the cost of stationery, computer consumables and computer time for sorting information. However

iv. price may be fixed in Indian currency as well as in Sterling Pound and American Dollar. Foreign currency prices may be determined using relevant official multiplier fixed from time to time for printed government publications;

v. Survey results/data should be made available to the data users in India and abroad simultaneously after the expiry of three years from the completion of the field work or after the reports based on survey data are released, whichever is earlier;

vi. Data users will give an undertaking in the prescribed form to the effect, inter alia, that the official statistics obtained by him for his own declared use will not be passed on with or without profit to any other data user or disseminator of data with or without commercial purpose;

vii. Data users will have to acknowledge the data sources in their research work based on official statistics. One copy of research study along with short summary of conclusions, if required by the concerned data source agency, should be supplied in the form of hard copy or on electronic media, free of cost; and

viii. The Department of Statistics will be the nodal agency for dissemination of official statistics provided by Central Government Ministries and Departments. However the concerned subject matter Ministries and Departments of the Central Government will be the final authority on issues arising out of this policy with a view to resolving any dispute between a data user and a data source agency.

The guidelines for the release of the data are:

i. A data warehouse in the Department of Statistics will be created to enable the data users and general public to have easy access to the published as well as unpublished validated data from one source.

ii. The data warehouse will collect data from various source agencies, integrate the data into logical subject areas, store the data in a manner that is accessible and understandable to non-technical decision-makers and deliver data/information to decision makers through report writing and query tools.

iii. As data source agencies are generating data at various levels, the responsibility of data supplied and receipt will be shared between the respective Central Ministries/Departments/ Agencies and the Department of Statistics by establishing and maintaining close collaboration.

iv. For each data type and source, detailed studies will be undertaken by the Department of Statistics in cooperation with the concerned data source agency on

(a) the concepts, definitions, classifications and methods used in data collection and processing including validation,
(b) formats of data collection,
(c) media on which data will be supplied,
(d) frequency of supply of data and
(e) procedures and modalities for preservation, updation and dissemination of data.

v. The volume of data flowing from each source agency into the data warehouse will be assessed by the Department of Statistics in order to formulate the various parameters required for designing, establishing and maintaining a data warehouse.

vi. Each data source agency will be required to adopt for itself a calendar for preparation and release of data data which it will share with the Department of Statistics. As part of its nodal responsibility of dissemination of data from the source the Department of Statistics will keep track of the data release calendar of each source agency.

vii. The data source agency will be required to supply on computer compatible media, validated data, published or unpublished free of cost to the data warehouse.

viii. The Department of Statistics will prepare Directories of all available data in the data warehouse and update the same at frequent intervals. A web site will be created for the data warehouse and Directories will be available on the web site.

ix. From the data warehouse, data/information will be made available free of cost to the data source agencies for official use and also the approved research institutes and universities for research purposes.

x. The price of data to be supplied to other users will depend upon system hardware and software used for data storage, retrieval etc. and also on medium of supply of data.

Annex 2

National Statistical Commission

The Government of India has setup a National Statistical Commission to critically examine the deficiencies of the present statistical system with a view to recommending measures for a systematic revamping of the system (Gazette of India, Extraordinary Part 1, No.10, Resolution No.M-13011/3/99-Admn.IV dated 19.01.2000). The Commission consists of Dr. C.Rangarajan, Governor, Andhra Pradesh, as its part-time Chairman and the following 11 eminent experts as part-time members:

1. Mr. V.R.Rao, ex-Director, Central Statistical Organisation and UN Advisor

2. Mr. S.M.Vidhwans, ex-Director (Economics & Statistics), Govt. of Maharashtra and UN expert

3. Prof. J.Roy, Professor Emeritus, Indian Statistical Institute

4. Dr. Prem Narain, Emeritus Scientist, IARI and ex-Director, Indian Agricultural Statistics Research Institute

5. Dr. Rakesh Mohan, Director-General, National Council of Applied Economic Research (NCAER)

6. Dr. V.R.Panchmukhi, Director-General, Research and Information System for the Non-Aligned and other Developing Countries

7. Dr. Y.Venugopal Reddy, Deputy Governor, Reserve Bank of India

8. Dr. K.Srinivasan, Executive Director, Population Foundation of India and ex-Director of International Institute of Population Studies

9. Prof S.Tendulkar, Delhi School of Economics and Vice-Chairman, N.A.B.S.

10. Dr. A.B.L.Srivastava, Chief Consultant, Educational Consultants India Ltd. & ex-Professor,National Council for Educational Research and Training (NCERT)

11. Dr. Fredie Ardeshir Mehta, Eminent private sector economist and Director, M/s Tata Sons Ltd.

The terms of reference of the National Statistical Commission are as follows:

1. To examine critically the deficiencies of the present statistical system in terms of timeliness, reliability and adequacy

2. To recommend measures to correct the deficiencies and revamp the statistical system to generate timely and reliable statistics for the purpose of policy and planning in Government at different levels of administrative structure

3. To recommend permanent and effective coordinating mechanism for ensuring integrated development of the decentralised statistical system in the country

4. To review the existing legislation for the collection of statistical information and to recommend amendments where necessary, to achieve the objective of collection and dissemination of timely, reliable and adequate statistics

5. To review the existing organisation of the Ministry of Statistics and Programme Implementation (Statistics Wing) and other statistical units of the Government and to make recommendations on their staffing and training requirements to enable them to cope with the increase and development of statistical sources

6. To examine the need for instituting statistical audit of the range of services provided by the Government and the local bodies and make suitable recommendations thereof and

7. To recommend any other measures for improving the statistical system in the country.

The Commission is expected to submit its report to the Government within a period of twelve months from the date of its establishment.

Annex 3

List of Special URLs for frequently accessed documents of Reserve Bank of India (RBI).

* Currency Museum http://www.museum.rbi.org.in
* Exchange Control Manual http://www.ecm.rbi.org.in
* Weekly Statistical Supplement http://www.wss.rbi.org.in
* RBI Bulletin http://www.bulletin.rbi.org.in
* Monetary and Credit policy http://www.cpolicy.rbi.org.in
* 9% Government of India Relief Bonds http://www.goirb.rbi.org.in
* RBI Notifications http://www.notifics.rbi.org.in
* RBI Press Releases http://www.pr.rbi.org.in
* RBI Speeches http://www.speeches.rbi.org.in
* RBI Annual Report http://www.annualreport.rbi.org.in
* Credit Information Review http://www.cir.rbi.org.in
* Report on Trend and Progress of Banking in India http://www.bankreport.rbi.org.in
* FAQs http://www.faqs.rbi.org.in
* Committee reports http://www.reports.rbi.org.in
* Y2K http://www.y2k.rbi.org.in
* FII List http://www.fiilist.rbi.org.in
* Electronics Clearing Service http://www.ecs.rbi.org.in
* Facilities for NRIs http://www.nri.rbi.org.in
* SDDS-National Summary data page-India http://www.nsdp.rbi.org.in
* Foreign Exchange Management Act, 1999 http://www.fema.rbi.org.in

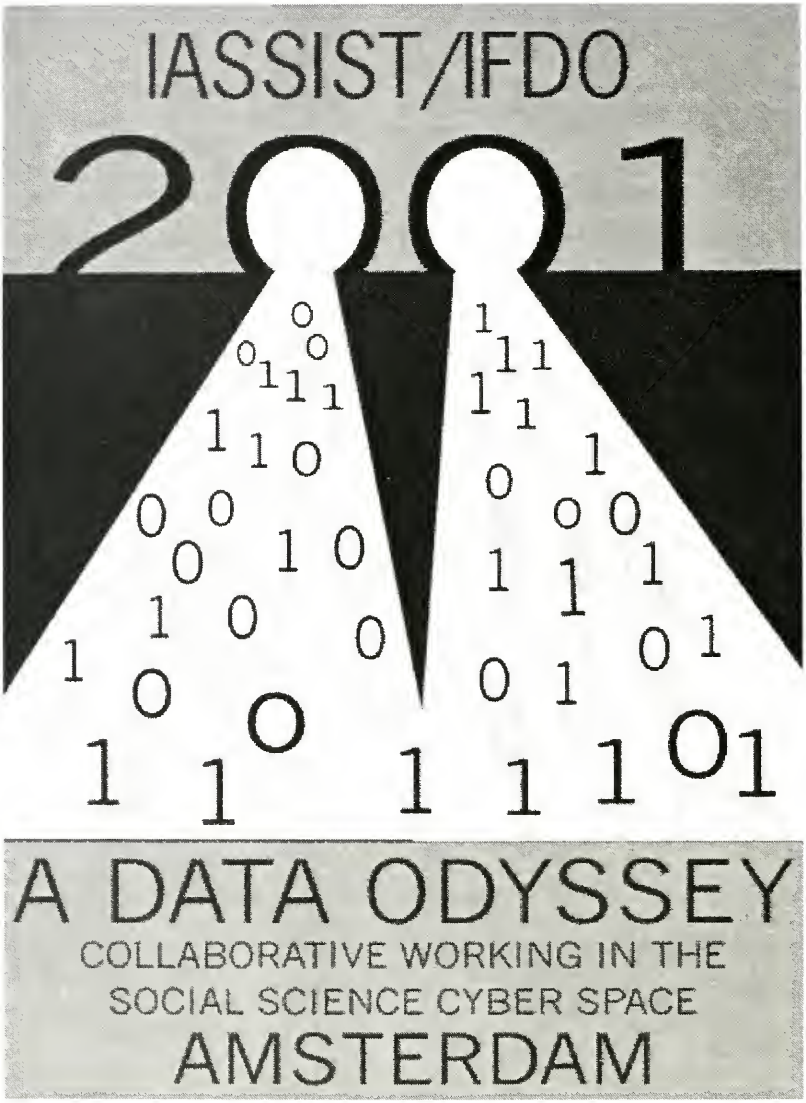
Annex 4

Reserve Bank of India.Weekly Statistical Supplement
Dec 02, 2000

1. Reserve Bank of India
2. Foreign Exchange Reserves
3. Scheduled Commercial Banks - Business in India
4. Interest Rates (per cent per annum)
5. Accommodation Provided by Scheduled Commercial Banks to Commercial Sector in the form of Bank Credit and Investments in Shares/Debentures/Bonds/Commercial Paper etc.

6. Foreign Exchange Rates - Spot and Forward Premia
7. Money Stock: Components and Sources
8. Reserve Money: Components and Sources
9. Auctions of 14-Day Government of India Treasury Bills
10. Auctions of 91-DayGovernment of India Treasury Bills
11. Auctions of 182-Day Government of India Treasury Bills
12. Auctions of 364-Day Government of India Treasury Bills
13. Certificates of Deposit Issued by Scheduled Commercial Banks
14. Commercial Paper Issued by Companies (At face value)
15. Index Numbers of Wholesale Prices (Base: 1993-94 = 100)
16. BSE Sensitive Index and NSE Nifty Index of Ordinary Share Prices - Mumbai
17a. Average Daily Turnover in Call Money Market
17b. Turnover in Government Securities Market (Face Value)
17c. Turnover in Foreign Exchange Market
17d. Weekly Traded Volume in Corporate Debt at NSE
18. Bullion Prices (Spot)
19. Government of India: Treasury Bills Outstanding (Face Value)
20. Government of India: Long and Medium Term Borrowings - 2000-2001
21. Secondary Market Transactions in Government Securities (Face Value)

Dr. Jagtar Singh and H. P. S. Kalra, jagtar@pbi.ernet.in harry@pbi.ernet.in, Department of Library & Information Science, Punjabi University, Patiala - 147 002. (India)


IASSIST/IFDO
2001
A DATA ODYSSEY
COLLABORATIVE WORKING IN THE
SOCIAL SCIENCE CYBER SPACE
AMSTERDAM

Collaborative Working in the Social Science CyberSpace

The International Association for Social Science Information Services and Technology (IASSIST) will hold its 27th annual conference with the International Federation of Data Organizations (IFDO) from May 14 - 19, 2001.

The conference will be convened in Amsterdam, capital of The Netherlands.

This year's conference emphasizes the need for co-operation on technical and organizational matters, and of course on contents.

IASSIST conferences bring together data professionals, data producers, and data analysts from around the world who are engaged in the creation, acquisition, processing, maintenance, distribution, preservation, and use of numeric social science data for research and instruction.

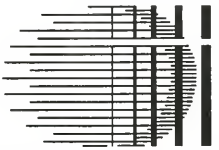
IFDO was established in 1977 in response to advanced research needs of the international social science community. IFDO stimulates to co-ordinate worldwide data services and thus enhance social science research.

Our conference title makes reference to a mythical story of a journey rich in challenge, danger and reward. Unlike those in the epic Odyssey, we cannot call on the help of gods, we have to solve the problems ourselves. IASSIST is the professional body that exists to foster co-operation among 'data workers' in their quest for data, connecting those who seek data with those who produce data through sharing between data archives and data libraries.

During the last quarter of the 20th Century, IASSIST has held its annual international conference in Europe about every four years. IASSIST will do so again, this time in Amsterdam in collaboration with IFDO. Join us in millennium mood, eager to chart the way through the mysteries of the virtual environment, assisting those who seek to discover and locate data, for research or teaching, to reap their just reward.

http://ww...

niwi WSA



IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Services and Technology (IASSIST)** is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:

Regular Membership. $40.00 per calendar year.

Student Membership: $20.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:

$70.00 per calendar year (includes one volume of the

Membership form

I would like to become a member of IASSIST. Please see my choice below:
Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:
http://datalib.library.ualberta.ca/iassist/mbrship2.html

- ☐ $50 (US) Regular Member
- ☐ $25 Student Member
- ☐ $70 Subscription (payment must be made in US$)
- ☐ List me in the membership directory
- ☐ Add me to the IASSIST listserv

Please make checks payable, *in US funds*, to IASSIST and Mail to:

IASSIST,
Assistant Treasurer
JoAnn Dionne
50360 Warren Road
Canton, MI 48187
USA

Name: _____

Job Title: _____

Organization: _____

Address: _____

_____

_____

City: _____ State/Province: _____

Postal Code: _____ Country: _____

Phone: _____ FAX: _____

E-mail: _____ URL: _____

Return Undelivered Mail To:

IASSIST QUARTERLY
c/o Wendy Treadwell
1758 Pascal St. North
Falcon Heights, MN 55113
USA

Serials Department(SERLIBS82186344)
Univ of North Carolina-Chapel Hill
CB #3938 Davis Library
Chapel Hill NC 27514-8890
U.S.A.